ITI8700: Knowledge Representation

05. Large Knowledge Bases: Overview

Martin Verrev

Spring 2024

Wordnet

A lexical database of semantic relations between words that links words into semantic relations including synonyms, hyponyms, and meronyms. The synonyms are grouped into **synsets** with short definitions and usage examples. It can thus be seen as a combination and extension of a dictionary and thesaurus.

https://wordnet.princeton.edu

See also: https://www.nltk.org/nltk_data/

DBPedia

A crowd-sourced community effort to extract structured content from the information created in various Wikimedia projects. This structured information resembles an open knowledge graph that is available for everyone on the Web.

https://www.dbpedia.org/

Wikidata

A collaboratively edited multilingual knowledge graph hosted by the Wikimedia Foundation. It is a common source of open data that Wikimedia projects such as Wikipedia and anyone else, is able to use. Wikidata is a wiki powered by the software MediaWiki, including its extension for semi-structured data, the Wikibase.

https://www.wikidata.org

Yago

YAGO (Yet Another Great Ontology) is an open source knowledge base developed at the Max Planck Institute for Informatics in Saarbrücken. It is automatically extracted from Wikipedia and other sources. The information in YAGO is extracted from Wikipedia (e.g., categories, redirects, infoboxes), WordNet (e.g., synsets, hyponymy), and GeoNames. The accuracy of YAGO was manually evaluated to be above 95% on a sample of facts. To integrate it to the linked data cloud, YAGO has been linked to the DBpedia ontology and to the SUMO ontology

https://yago-knowledge.org/

BabelNet

A multilingual lexicalized semantic network and ontology developed at the NLP group of the Sapienza University of Rome. BabelNet was automatically created by linking Wikipedia to the most popular computational lexicon of the English language, WordNet. The integration is done using an automatic mapping and by filling in lexical gaps in resource-poor languages by using statistical machine translation.

https://babelnet.org

ConceptNet

A semantic network based on the information in the OMCS database. ConceptNet is expressed as a directed graph whose nodes are concepts, and whose edges are assertions of common sense about these concepts. Concepts represent sets of closely related natural language phrases, which could be noun phrases, verb phrases, adjective phrases, or clauses

https://conceptnet.io

NELL

Never-Ending Language Learning system (NELL) is a semantic machine learning system that as of 2010 was being developed by a research team at Carnegie Mellon University. NELL was programmed by its developers to be able to identify a basic set of fundamental semantic relationships between a few hundred predefined categories of data, such as cities, companies, emotions and sports teams.

http://rtw.ml.cmu.edu

CYC

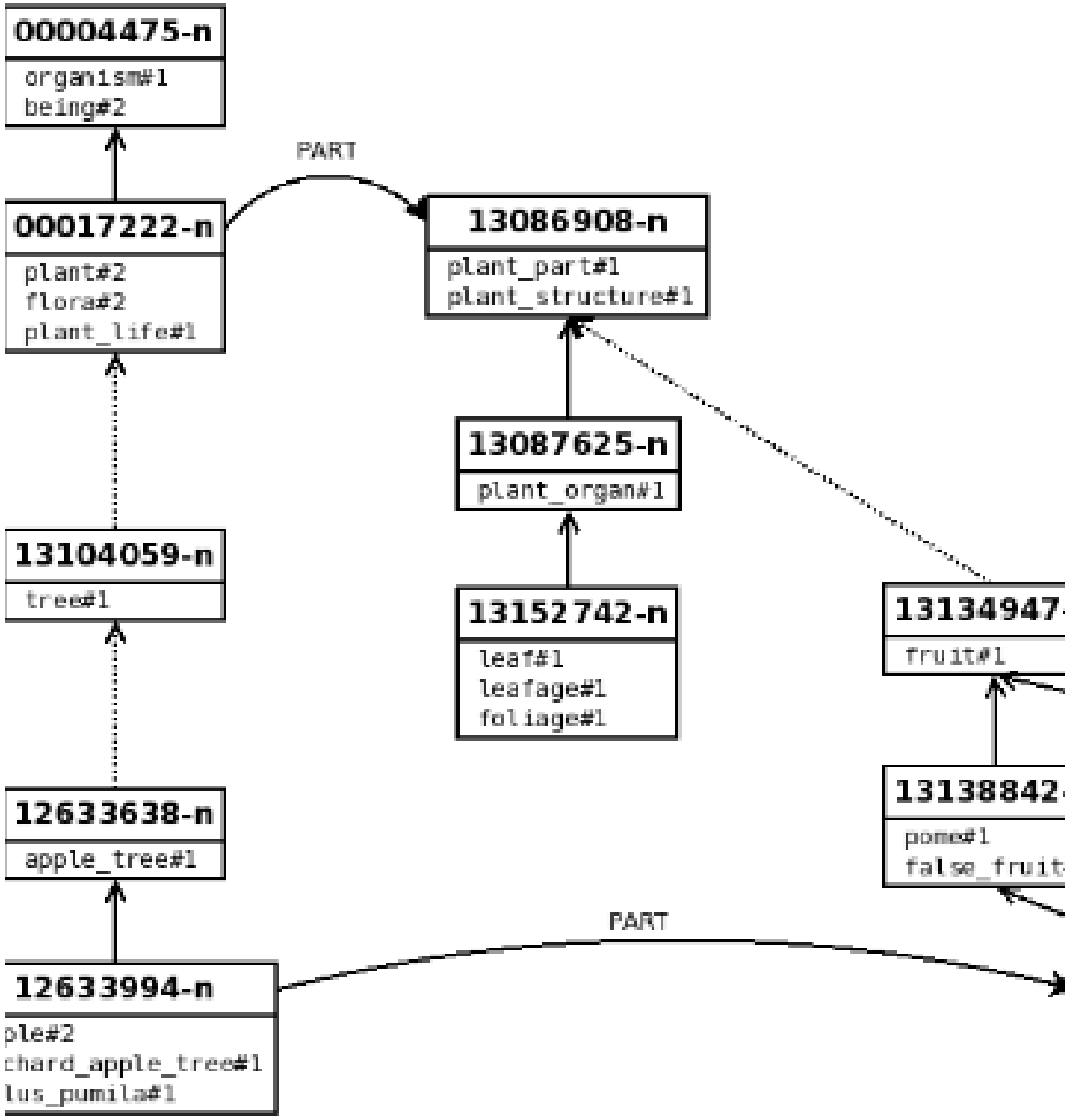
A long-term artificial intelligence project that aims to assemble a comprehensive ontology and knowledge base that spans the basic concepts and rules about how the world works. Hoping to capture common sense knowledge, Cyc focuses on implicit knowledge that other AI platforms may take for granted.

https://cyc.com/

Adimen-SUMO

An off-the-shelf first-order ontology that has been obtained by reengineering out of the 88% of SUMO (Suggested Upper Merged Ontology). Adimen-SUMO can be used appropriately by FO theorem provers for formal reasoning.

https://adimen.si.ehu.es/web/adimen SUMO



TPTP

TPTP (Thousands of Problems for Theorem Provers) is a freely available collection of problems for automated theorem proving. It is used to evaluate the efficacy of automated reasoning algorithms. The roblems are expressed in a simple text-based format for first order logic or higher-order logic. TPTP is used as the source of some problems in CASC.

https://tptp.cs.miami.edu/

TPTP Axioms

- Wordnet: http://tptp.cs.miami.edu/cgi-bin/SeeTPTP?Category=Axioms&File=NLP001+0.ax
- CYC: http://tptp.cs.miami.edu/cgi-bin/SeeTPTP?Category=Axioms&File=CSR002+1.ax
- SUMO: http://tptp.cs.miami.edu/cgi-bin/SeeTPTP?Category=Axioms&File=CSR003+0.ax

Schema.org

Schema.org is a reference website that publishes documentation and guidelines for using structured data mark-up on web-pages (called microdata). Its main objective is to standardize HTML tags to be used by webmasters for creating rich results (displayed as visual data or infographic tables on search engine results) about a certain topic of interest. It is a part of the semantic web project, which aims to make document mark-up codes more readable and meaningful to both humans and machines.

https://schema.org/

WolframAlpha

An answer engine developed by Wolfram Research.[3] It is offered as an online service that answers factual queries by computing answers from externally sourced data.

https://reference.wolfram.com/language/guide/KnowledgeRepresentationAndAccess.html

FrameNet

A group of online lexical databases based upon the theory of meaning known as Frame semantics, developed by linguist Charles J. Fillmore. The project's fundamental notion is simple: most words' meanings may be best understood in terms of a semantic frame, which is a description of a certain kind of event, connection, or item and its actors.

http://framenet.icsi.berkeley.edu/

PropBank

PropBank is a corpus that is annotated with verbal propositions and their arguments—a "proposition bank". Although "PropBank" refers to a specific corpus produced by Martha Palmer, the term propbank is also coming to be used as a common noun referring to any corpus that has been annotated with propositions and their arguments.

https://propbank.github.io/

ARG0	agent	ARG3	starting point, benefactive, attribute
ARG1	patient	ARG4	ending point
ARG2	instrument, benefactive, attribute	ARGM	modifier

Table 1.1: List of arguments in PropBank

Other Frame Datasets

VerbNet is a project that maps PropBank verb types to their corresponding Levin classes (https://verbs.colorado.edu/verbnet/). See also: https://verbnetparser.com/

SemLink is a project that brings together different lexical resources: PropBank, VerbNet, FrameNet, WordNet (https://verbs.colorado.edu/semlink/)

Unified Verb Index is a system which merges links and web pages from four different natural language processing projects: VerbNet, PropBank, FrameNet, Ontonotes (https://verbs.colorado.edu/verb-index/)

Allen-AI

- **Aristo TupleKB.** Dataset contains a collection of high-precision, domain-targeted (subject,relation,object) tuples extracted from text using a high-precision extraction pipeline, and guided by domain vocabulary constraints.
  https://allenai.org/data/tuple-kb
- **ATOMIC: An Atlas of Machine Commonsense** for If-Then Reasoning is a large-scale common sense repository of textual descriptions that encode both the social and the physical aspects of common human everyday experiences, collected with the aim of being complementary to commonsense knowledge encoded in current language models.
  https://allenai.org/data/atomic, demo:
  https://mosaickg.apps.allenai.org/kg_atomic2020

Max Planck Institute

- **Ascent++** is a commonsense knowledge base (CSKB) constructed from the cleaned Common Crawl data. It consists of 2 million CSK assertions about 10K popular concepts, presented in the established ConceptNet schema with 19 predicates (e.g., AtLocation, CapableOf, HasProperty, etc.). As of 2022, it presents the highest-quality automated CSKB, both in terms of precision, and in terms of ranked recall. https://ascentpp.mpi-inf.mpg.de/
- **Quasimodo** is a commonsense knowledge base that focuses on salient properties of objects. https://quasimodo.mpi-inf.mpg.de/

Max Planck Institute

- **WebChild** is a large collection of commonsense knowledge, automatically extracted and disambiguated from Web contents. It contains triples that connect nouns with adjectives via fine-grained relations. The arguments of these assertions, nouns and adjectives, are disambiguated by mapping them onto their proper WordNet senses. https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/commonsense/webchild
- **UncommonSense** is a framework for materializing informative negative commonsense statements. Given a target concept, comparable concepts are identified in the CSKB, for which a local closed-world assumption is postulated. This way, absent positive statements about comparable conceptbecome seeds for negative statement candidates. The large set of candidates is then scrutinized, pruned and ranked by informativeness. https://uncommonsense.mpi-inf.mpg.de/

Thank you! :)